2010 Demonstration Privacy-Protected Microdata Files 2021-04-28

Over the past several months, the Census Bureau has been actively tuning the parameters of the 2020 Census Disclosure Avoidance System (DAS) to ensure fitness-for-use of the P.L. 94-171 Redistricting data product for the redistricting and Voting Rights Act use cases. Over the past eighteen months, our development and tuning of the DAS has benefited substantially from feedback from our federal advisory committees, stakeholder groups, and our data users and from the continuing support of the Committee on National Statistics' (CNSTAT) expert group. To enable this invaluable feedback, we have released a series of demonstration data products using 2010 Census data for evaluation. In early June 2021, the Data Stewardship Executive Policy (DSEP) Committee will be making final decisions on the DAS parameters to be used for production of the 2020 Census redistricting data. Before those decisions are made, we would like to provide the public with another opportunity to evaluate our DAS implementation and to provide feedback that can inform our final decision-making for the redistricting data product.

To facilitate that feedback, we are pleased to announce the release of additional demonstration data, generated by running 2010 Census data through the DAS. These demonstration data contain Detailed Summary Metrics (DSM) and Privacy-Protected Microdata Files (PPMFs) for two complete runs of the 2010 Census redistricting data, at different levels of privacy-loss budget.

Detailed Summary Metrics

The Detailed Summary Metrics we released for these DAS data runs allow our data users to assess these improvements and their impact on fitness-for-use in a variety of ways. They provide a variety of accuracy measures for a range of use cases that our data users have identified. Taken together, the detailed summary metrics provide a comprehensive snapshot of the overall fitness-for-use of the resulting data. That said, we recognize that our data users assess accuracy and fitness-for-use for diverse use cases in very different ways, so we are also releasing Privacy-Protected Microdata Files for users to perform more specific analyses that reflect their particular use cases.

Privacy-Protected Microdata Files

Privacy-Protected Microdata Files (PPMFs) are the underlying microdata files for the entire nation used to generate the Detailed Summary Metrics. It is important to note that while the data in the PPMFs look like individual records, all of the data are privacy-protected. The microdata records generated by the DAS ensure respondent privacy through the application of differentially private statistical noise. The microdata included in the PPMF do not include any

actual census responses. They are simply the microdata format, generated by the DAS, and used by the Census Bureau's tabulation production system to produce privacy-protected tables.

While these PPMFs are untabulated microdata records, the IPUMS National Historic Geographic Information System (NHGIS) will be tabulating, formatting and posting data tables for direct comparison to published 2010 Census tabulations. This partnership allows the census staff who would otherwise perform the time-intensive tabulation, data review and release process in-house to continue their focus on other important data collection and processing work.

Privacy-Loss Budgets

The Census Bureau released the first set of demonstration data products for our data users to evaluate in October 2019. Over the subsequent year, we released additional sets of demonstration data (in May 2020, September 2020, and November 2020) to allow our data users to review and assess improvements to the DAS algorithms. Throughout this process, however, we maintained the conservative PLB set for the initial demonstration data product. While we recognize that this decision to hold the PLB constant across the demonstration runs meant that the resulting data would have substantially more noise (error) than should be expected in the final 2020 Census data products, holding the PLB constant enabled us and our data users to home in on the elements of the algorithm that were causing systemic distortions that needed to be addressed. We acknowledge that this has unfortunately led some of our data users to expect comparable amounts of noise in the final 2020 Census data.

The April 28, 2021 demonstration data feature a higher PLB, which more closely approximates the anticipated level that DSEP will set for the final 2020 Census redistricting data product. This higher PLB tunes the resulting data for greater accuracy and ensures that they meet the accuracy targets that we have established for redistricting, Voting Rights Act enforcement, and other priority uses of the redistricting data. As our tuning of the DAS reflects additional improvements made since our last demonstration data release in November 2020, we recognize that some of our data users may wish to evaluate these improvements against the November 2020 (and earlier) demonstration files independent of the increased PLB. Consequently, in addition to the demonstration data reflecting the anticipated, higher PLB, we are also releasing a version of the demonstration data that maintains the lower PLB used for the prior releases.

The six files included in this release are:

Global PLB of ε=12.2 (tuned to accuracy targets)

- Detailed Summary Metrics
- Person-level data (ε=10.3)
- Unit-level data (ε=1.9)

Global PLB of ε=4.5 (for algorithmic comparison to prior demonstration data releases)

- Detailed Summary Metrics
- Person-level data (ε=4.0)
- Unit-level data (ε=0.5)

For More Information, see: Developing the DAS: Progress Metrics and Data Runs Web Page

Improvements and Tuning Reflected in This Release

As we discussed in our April 7, 2021 newsletter, the parameters of the TDA can be varied in a number of ways: query strategy, allocation of PLB across geographic levels, allocation of PLB across queries, and optimization of geographic post-processing to improve accuracy of the data for "off-spine" geographic entities. Determining the optimal settings for these parameters requires empirically evaluating large numbers of TDA runs against objective accuracy metrics.

Accuracy Targets

For the P.L. 94-171 redistricting data product, the principal statutory use cases are the redistricting process and the U.S. Department of Justice's enforcement of the Voting Rights Act of 1965 (VRA). To facilitate this analysis, the Department of Justice supplied sample redistricting and VRA use cases for the Census Bureau to evaluate against.

Based on these use cases and additional feedback, we created an accuracy target to ensure that the largest racial or ethnic group in any geographic entity with a total population of at least 500 people is accurate to within 5 percentage points of their enumerated value at least 95% of the time.

Because the redistricting and VRA use cases rely on geographic aggregations that cannot be prespecified (e.g., precincts and wards that will be drawn after the data are published), for evaluation purposes the DAS Team used three already specified geographic constructs that resemble the size and composition of voting districts that will eventually be drawn: block groups (which are on the TDA geographic spine), places (which are "off-spine"), and a customized set of off-spine entities that distinguished between strong minor civil division states and other states. The customized off-spine entities are similar to census designated places.

Because these accuracy targets are expressed in relative shares of the total population, tuning the TDA for accuracy of the racial/ethnic group's share also tunes the algorithm for corresponding accuracy of the total population of those geographies.

Query Strategy

The DAS TopDown Algorithm (TDA) operates by taking a series of measurements (queries) of the tabulations that support the redistricting data product, adding a small amount of uncertainty (noise) to each of those queries to protect privacy, then converting the results of those queries back into individual-level records for the entire population. These queries can be structured in a number of different ways, with implications for the relative accuracy of different sets of cross-tabulations by demographic characteristics.

The query strategy used for the April 28, 2021 demonstration data used the following queries for the person-level data:

TOTAL POPULATION
CENRACE *(all 63 allowed combinations of the OMB-designated race categories)*
HISPANIC *(Hispanic, not Hispanic)*
VOTINGAGE *(≥18 years, <18 years of age)*
HHINSTLEVELS *(institutional vs. non-institutional group quarters types)*
HHGQ *(household and group quarters types)*
HISPANIC*CENRACE
VOTINGAGE*CENRACE
VOTINGAGE*HISPANIC
VOTINGAGE*HISPANIC*CENRACE

DETAILED (HHGQ x VOTING_AGE x HISPANIC x CENRACE)

PLB Allocation

The relative accuracy of different tabulations similarly depends on the share of the PLB allocated to each of the queries performed by the algorithm. Queries for smaller tabulations or cross-tabulations, like total population counts or voting age population counts, can be very accurate for any geographic level even with minimal allocation of PLB. Queries for cross-tabulations with a large number of categories (e.g., VOTINGAGE*HISPANIC*CENRACE, with 252 different combinations) require larger allocations of PLB to achieve comparable levels of accuracy.

PLB allocation by query for the April 28, 2021 demonstration data was finely tuned at different levels of geography to meet the accuracy targets discussed above. In general, however, PLB was allocated proportionally by the size of the query, with the DETAILED query (HHGQ x VOTING_AGE x HISPANIC x CENRACE) receiving the largest share of PLB.

Additional allocations of PLB were made to particular queries at specific geographic levels to further enhance the accuracy of certain statistics. For example, extra PLB was allocated to the total population query at the Block Group level to improve population counts for many "off-spine" geographic entities like places.

Geographic Hierarchy

The April 28, 2021 demonstration data also incorporates a significant change to how the TDA conducts the post-processing of the noisy measurements (see our June 23, 2020 newsletter for an explanation of post-processing) to improve the accuracy of data for off-spine geographies.

The TDA's standard geographic hierarchy (spine) follows the traditional Census tabulation geographies: Nation, State, County, Tract, Block Group, Block. However, ensuring fitness-for-use for the redistricting, Voting Rights Act, and other important use cases requires that we meet accuracy targets for geographic areas (e.g., voting districts, Minor Civil Divisions, and places) that fall off of that tabulation geography hierarchy.

If we relied on the standard tabulation geographic hierarchy, accuracy of these data would deteriorate the farther these geographic areas get from those included on the spine. To improve accuracy for these off-spine areas, we implemented a dynamic optimization strategy

to bring these areas closer to the post-processing geographic hierarchy. This *optimized spine* also separates out the post-processing for group quarters facilities at the Block Group level, to ensure that the characteristics of group quarters residents do not diffuse into their surrounding block group populations, or vice versa. It is important to note that these changes are exclusively used during the post-processing stage of the TDA; the resulting data will still be tabulated for release using the traditional tabulation geographies.

The optimized spine used for the April 28, 2021 demonstration data also maintains the separation of the American Indian and Alaska Native Tribal Area TDA post-processing geographic hierarchy that we introduced for the September 2020 and November 2020 demonstration data.

Data User Feedback

We look forward to feedback from data users on this new demonstration product. Please examine the new Detailed Summary Metrics and PPMFs at the higher global PLB of ε=12.2. Your feedback will inform our early June 2021 final decision on the PLB and on the 2020 Census redistricting data parameters. The deadline to submit feedback is **May 28, 2021**. Please send comments to 2020DAS@census.gov with the subject line "April 2021 Demonstration Data." Particularly useful feedback would describe:

- **Fitness-for-use:** Based on your analysis, would the data needed for your applications (estimates, projections, funding data sets, etc.) be satisfactory?
    - How did you come to that conclusion?
    - If your analysis found the data to be unsatisfactory, how incrementally would accuracy need to change to improve the use of the data for your required or programmatic use case(s)?
    - Have you identified any improbable results in the data that would be helpful for us to understand?"
- **Privacy:** Do the proposed products present any confidentiality concerns that we should address in the DAS?
- **Improvements:** Are there improvements you've identified that you want to make sure we retain in the final design? Be specific about the geography and error metric for the proposed improvement.